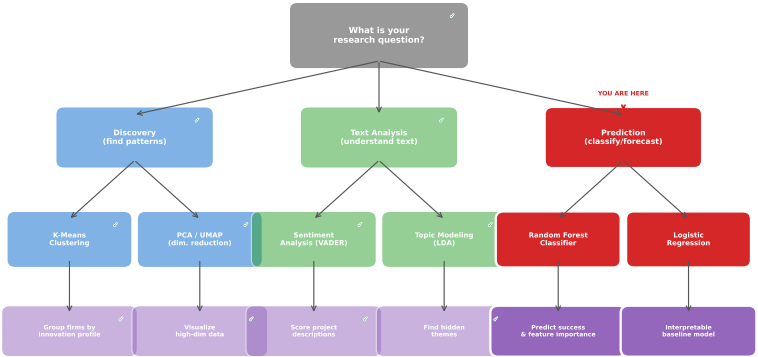
ML Technique Decision Tree



Your research question determines your method. We are exploring the Prediction branch.

Random Forest: Wisdom of the Crowd

The analogy:

Imagine 100 decision trees, each looking at a random subset of features. They each make a judgment. The majority vote is the prediction.

How it works:

1. Build 100+ decision trees
2. Each tree uses random features
3. Each tree trained on random samples
4. Final prediction = majority vote

Why Random Forest for research:

- Handles mixed feature types
- Robust to outliers
- Built-in feature importance
- Works well with small-medium datasets
- Hard to "break" (few hyperparameters)

Research-friendly

Feature importance ranks your variables by predictive power — directly publishable.

Random Forest = wisdom of crowds for machines. Reduces variance while maintaining low bias.

Logistic Regression: The Interpretable Baseline

You may already know this one:

Linear model for binary outcomes.
$P(\text{success}) = \sigma(\beta_0 + \beta_1 x_1 + \ldots)$

Advantages:

- Coefficients are interpretable
- Well-understood in management research
- Confidence intervals and p-values
- Good baseline for comparison

Compare with Random Forest:

- If RF barely beats logistic $\rightarrow$ relationships are mostly linear
- If RF significantly better $\rightarrow$ non-linear patterns exist
- Report both in your paper

DBA tip

Clean, transparent, interpretable. Your thesis committee loves this one. Use RF for discovery, logistic for confirmatory analysis.

$\sigma(z) = \frac{1}{1+e^{-z}}$ **(sigmoid function). "For every unit increase in R&D spending, the odds of success increase by X%."**

Train/Test Split: Why You Can't Grade Your Own Exam

The problem:

You can't test your theory on the same evidence you used to build it. That's circular reasoning.

Solution: hold-out validation

1. Split data: 80% train, 20% test
2. Train only on 80%
3. Evaluate on the held-out 20%
4. That accuracy is your real performance

Cross-validation (better):

- Split into 5 equal folds
- Train on 4 folds, test on 1
- Repeat 5 times, each fold as test
- Report mean $\pm$ standard deviation

Critical for DBA

With small samples ($n < 500$), cross-validation is essential. A single random split can be misleading.

Overfitting = model memorizes noise. Hold-out validation catches this. Always report CV results, not training accuracy.

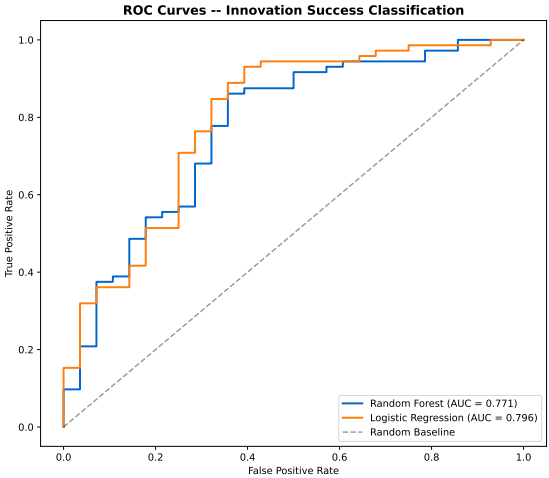
Live Demo: Predicting Innovation Success

Jupyter Notebook: 05_classification.ipynb

- Prepare features (encode categoricals)
- Train Random Forest + Logistic Regression
- 5-fold cross-validation
- Feature importance bar chart
- ROC curve and confusion matrix
- Partial dependence plots

Watch for: Which features drive success? Does the model agree with your Session 1 intuitions?

The ROC Curve: How Good Is Your Model?



What it shows:

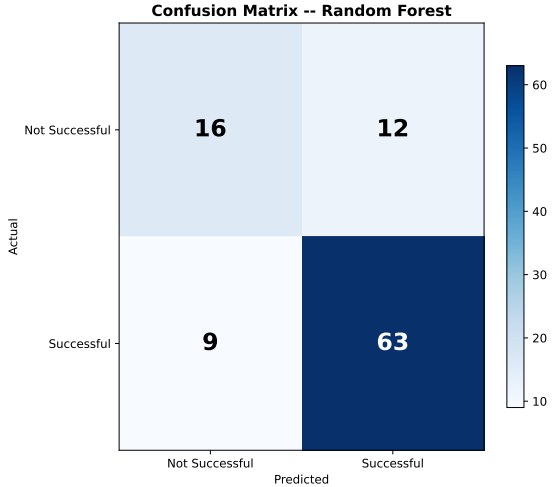
- X-axis: False positive rate
- Y-axis: True positive rate
- Diagonal = random guessing
- Upper-left corner = perfect

AUC = Area Under Curve

- 0.5 = useless
- 0.7–0.8 = acceptable
- 0.8–0.9 = good
- > 0.9 = excellent

At every threshold: how many true successes do you correctly identify vs. how many failures do you wrongly call successes?

Confusion Matrix: Types of Errors



Four outcomes:

- True Positive: Predicted success, was success
- True Negative: Predicted failure, was failure
- False Positive: Predicted success, was failure
- False Negative: Predicted failure, was success

For research:
Every model makes mistakes — the question is *what kind* of mistakes, and do they matter?

Precision = TP/(TP+FP). Recall = TP/(TP+FN). F1-score = harmonic mean of both.

Reporting for Publication

What to report:

1. Dataset description (N, features, labels)
2. Preprocessing steps taken
3. Algorithm(s) used + hyperparameters
4. Validation method (5-fold CV)
5. Performance metrics (accuracy, AUC, F1)
6. Feature importance ranking
7. Comparison with baseline (logistic)

Critical discussion:

- **Correlation $\neq$ causation**
  Feature importance shows *what predicts*, not *what causes*
- **Sample size limitations**
  500 projects is sufficient but not large
- **Generalizability**
  Results may differ for other countries/periods
- **Ethical considerations**
  Who benefits from these predictions?

Follow TRIPOD guidelines for reporting ML/prediction studies. Transparency builds credibility.

Your Turn: Where Does This Fit?

"Your model says R&D spending is the #1 predictor of success.
Your supervisor says 'correlation isn't causation.'
How do you respond?"

Take 60 seconds. Discuss with your neighbor.

Map Check: Where Are We?

Branches explored:

- ✓ Discovery + Text + Prediction branches complete
- ✓ Feature importance ranked
- ✓ Models validated with cross-validation

Next on the map:

- GenAI: a new tool beyond the tree
- Structured output for scalable coding

All three branches explored. Next: a tool that extends every branch — Generative AI.